Modelling and categorisation of seismic waves

Modelling and categorisation of seismic waves

Duncan McGregor[1], Andrew Bell[2], Bruce J. Worton[1]

[1] School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, Edinburgh, UK
[2] School of GeoSciences, The University of Edinburgh, Edinburgh, UK

E-mail for correspondence: `Bruce.Worton@ed.ac.uk`

Abstract: A key problem in seismology is assessing the similarity or difference between events, and constructing categorisations based on these measures. Clustering algorithms remain an active area of research, but many approaches are well documented. This paper assesses the suitability of a selection of these approaches to the problem of seismic waves, with reference to a data set taken from Tungurahua, Ecuador between 6–13 April 2015. In addition, approaches to modelling the waves, both parametric and nonparametric are fitted and assessed, and the suitability of certain data transformations are considered.

Keywords: Seismic waves, Clustering, Nonparametric.

1 Introduction

To better understand the processes occurring within volcanoes, seismologists study the seismic waves generated within — these are vibrations that propagate through the earth, and are measured by seismometers located near the volcano.
The raw seismic data is processed to identify distinct 'seismic events' where the seismic activity rises above some level. There are many aspects to the study of these events, however one key problem concerns their categorisation with the intention of identifying events that are similar and likely to have arisen from a common source. Seismologists can then use this categorisation to trace the evolution of the number of events occurring before, during, and after periods of activity, and make inferences about the nature of the events giving rise to activity.
This process of categorisation is currently not well-defined, with a variety of similarity measures and ad-hoc approaches to clustering in use. Plau-

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

sible approaches exist and are in wide use, but these are computationally expensive and have practical shortcomings we will discuss. The analysis is further complicated by the presence of ambient vibrations arising from natural and man-made sources, and any categorisation must allow for the presence of significant noise in the event signals.

2 Data set and modelling

We study modelling seismic data for 4805 events between 06/04/2015 and 13/04/2015 recorded at the station next to the Tungurahua volcano. Each event data item records the velocity of the vibration at 3001 distinct equally spaced points in time over a period of 30 seconds.

Our attempts to fit flexible parametric models to the data proved very challenging due to computational issues. An exhaustive search of suitable models would be impossible, but whilst the failure of our efforts cannot confirm the task is impossible. However, we have considered various possible nonparametric models (Hastie et al, 2009). All appear to give reasonable fits to the data, but wavelets appear to best capture the behaviour of the model (Donoho and Johnstone, 1994), and lend themselves readily to the smoothing of the wave (to remove ambient noise) and downsampling to reduce the dimension of the problem.

3 Results

Figure 1 gives an example of an event. Note the noisy nature of the data which leads to problems with parametric fits.

The gap statistic analysis provides useful summary information concerning the clusters. For example, on Day One there is no strong evidence for more than 6 groups. A typical example of a dendrogram (see Figure 2) for a day of events is rather crowded (making it difficult to identify individual events). However the high level structure is clearly visible and it is instructive to compare the distance of the different numbers of clusters with the Gap Statistic.

4 Conclusions

We started this work with the intention of investigating methods of categorising seismic waves. At the end, we have arrived at the following conclusions.

1. Seismic events are complex, and not readily modelled using a simple parametric approach.

2. Smoothing of events can be performed through a variety of nonparametric techniques, or by low-band pass.

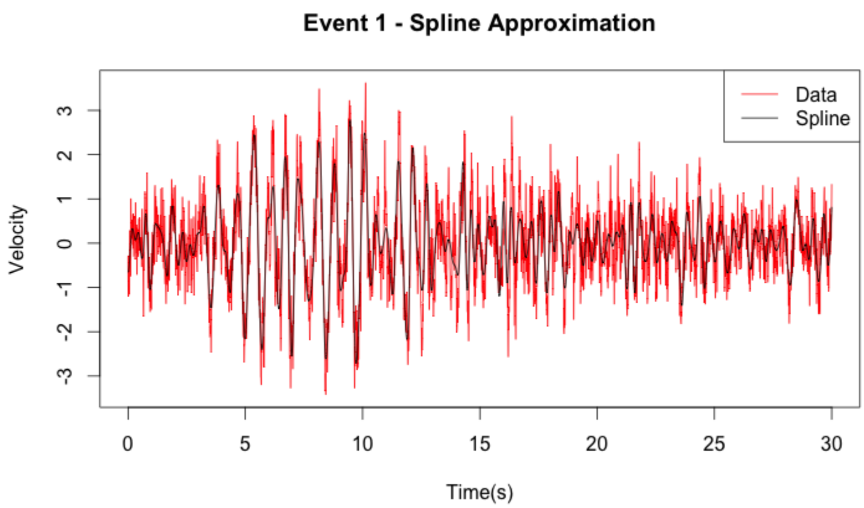

FIGURE 1. An event: example of the type of data being modelled.

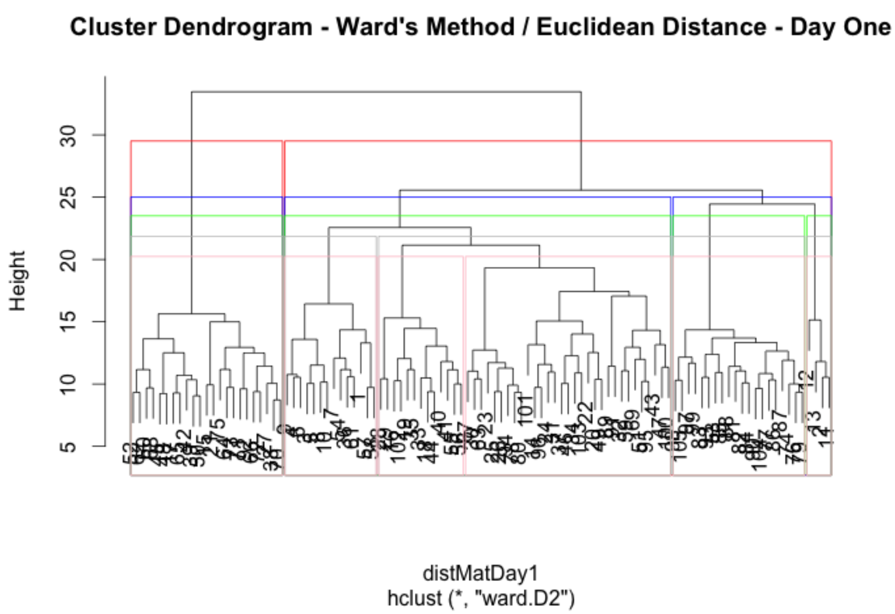

FIGURE 2. Spectral data, Euclidean distance, Ward's method on Day 1. This suggests 2 to 6 groups.

3. Seismic events data require transformation before they can be compared directly. Transformations include: (a) Scaling to match amplitude of events; (b) Translation to align events; (c) Windowing to eliminate regions where the signal to noise ratio is too low to be useful; (d) Smoothing to eliminate high frequency noise; (e) Fourier Transform to shift the event from the time domain to the frequency domain; (f) Principal Component Analysis to reduce the dimension of the problem.

4. Scaling and translation are essential for clustering techniques using conventional distance measures.

5. Windowing appear to stabilise the clustering under different clustering methods and distance measures.

6. Clustering methods give quite different groupings for smoothed signals. Relatedly, groupings are not invariant under decimation of the wavelet smoothed signal (Donoho and Johnstone, 1994).

7. Principal Component Analysis can reduce the dimension of our data from 3001 dimensions to 200 and retain over 80% of the observed variance.

8. The existing technique of cross-correlation looks sensible and fit for purpose, however we propose an alternative technique of carrying our Gap Analysis on the Spectral Intensity Data (Tibshirani et al, 2001). The brief simulation study we carried out suggests the technique requires refinement, and that spectral intensity data may not be the optimal choice, and applying Gap analysis on the raw data using cross correlation as a similarity measure may be a superior technique.

Acknowledgments: We are particularly grateful to The Instituto Geofisico of the Escuela Politecnica Nacional (IGEPN) of Ecuador for all their hard monitoring work and providing the data.

References

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425 – 455.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Second Edition. New York: Springer.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B*, **63**, 411 – 423.